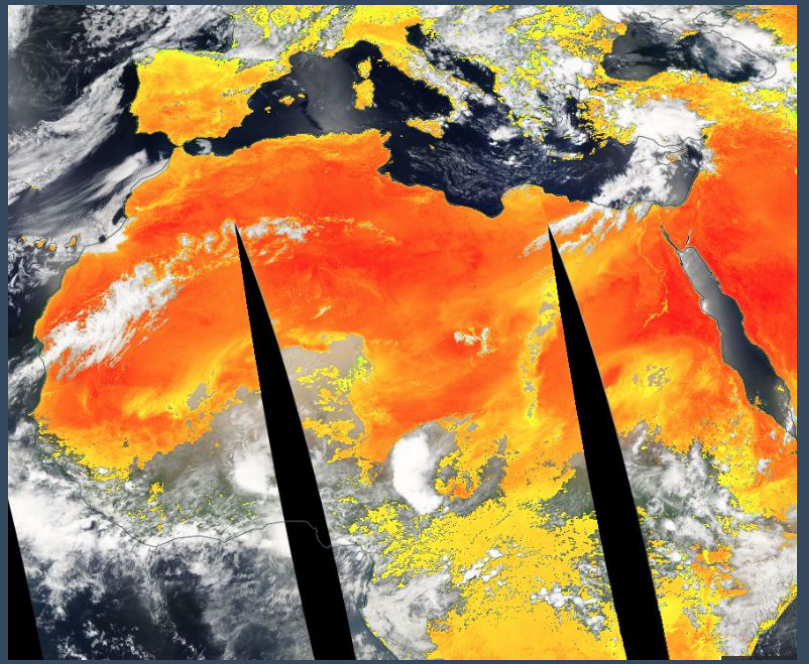

# Generalizing a Data Analysis Pipeline in the Cloud to Handle Diverse Use Cases in NASA's EOSDIS

*Christopher Lynnes*, NASA/Goddard Space Flight Center*
*Rahul Ramachandran*, NASA/Marshall Space Flight Center*

*Authors are both NASA Civil Servants

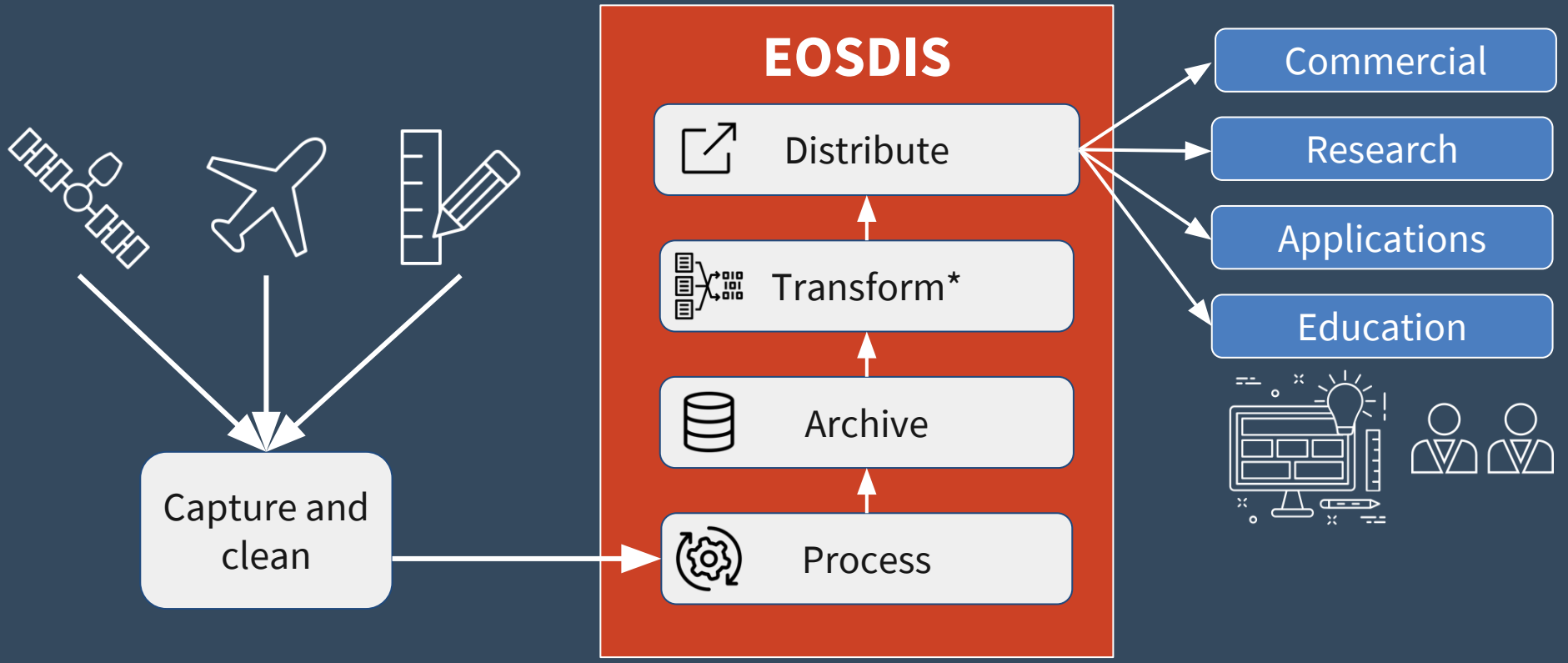
# NASA's Earth Science Data Systems Program

- Actively manages NASA's Earth science data as a national asset: satellite, airborne, and in situ
- Develops capabilities to support rigorous science research
- Processes instrument data to create high quality long-term Earth science data records.



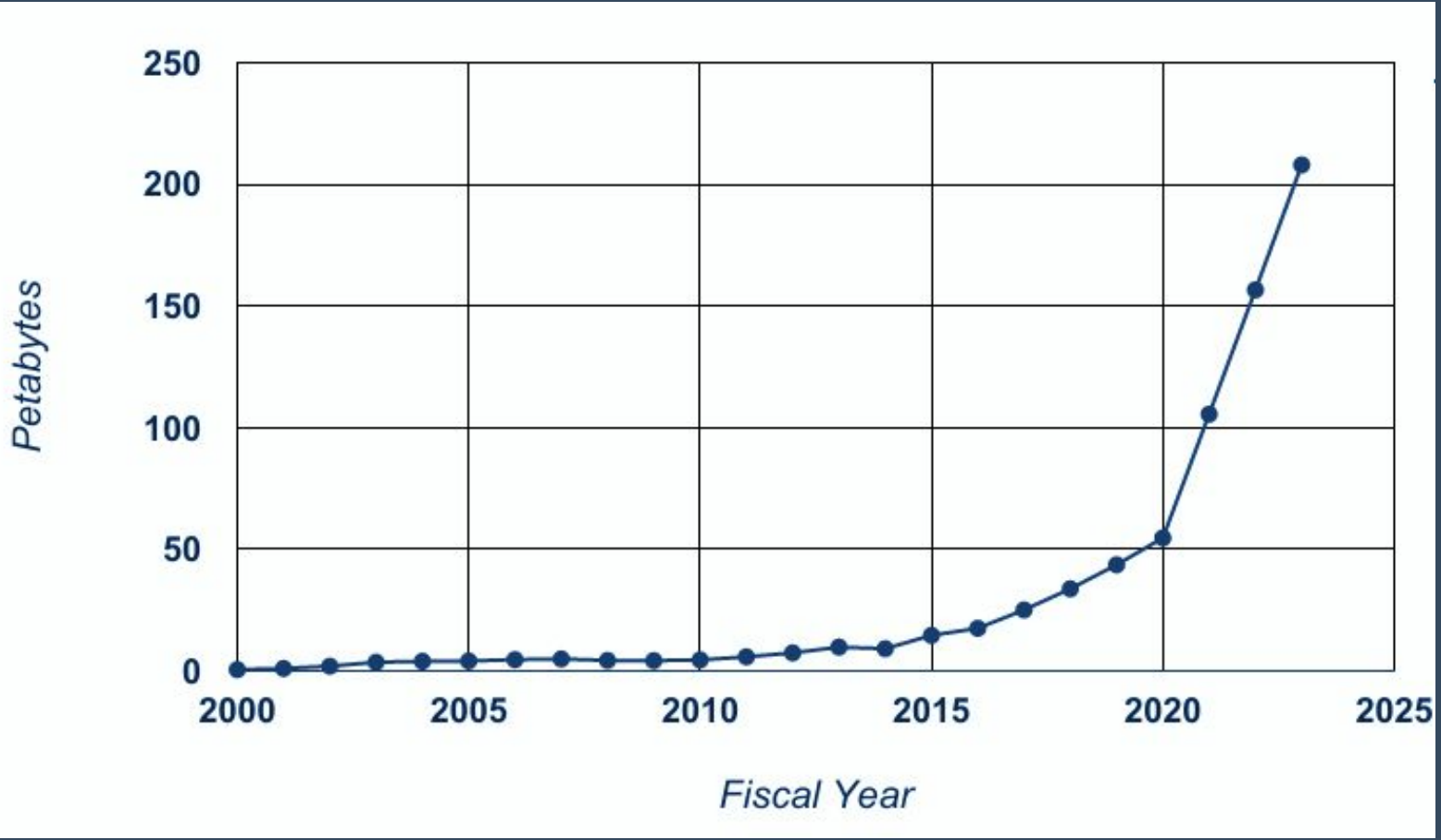
*Land Surface Temperature on a base of Corrected Reflectance from Aqua Moderate Resolution Imaging Spectroradiometer, 16 Jun 2018*

# Earth Observing System Data and Information System

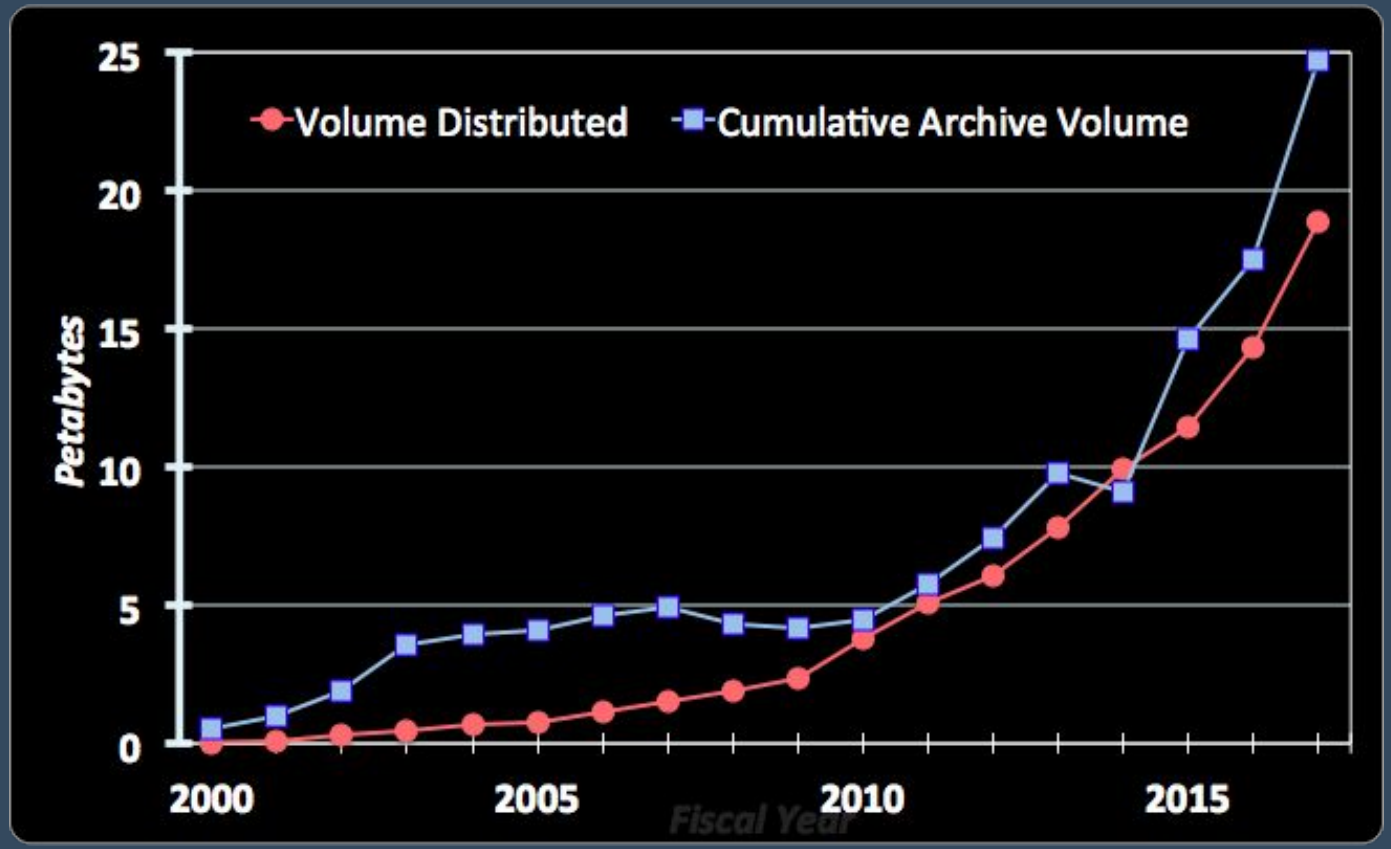



*Subset, reformat, reproject

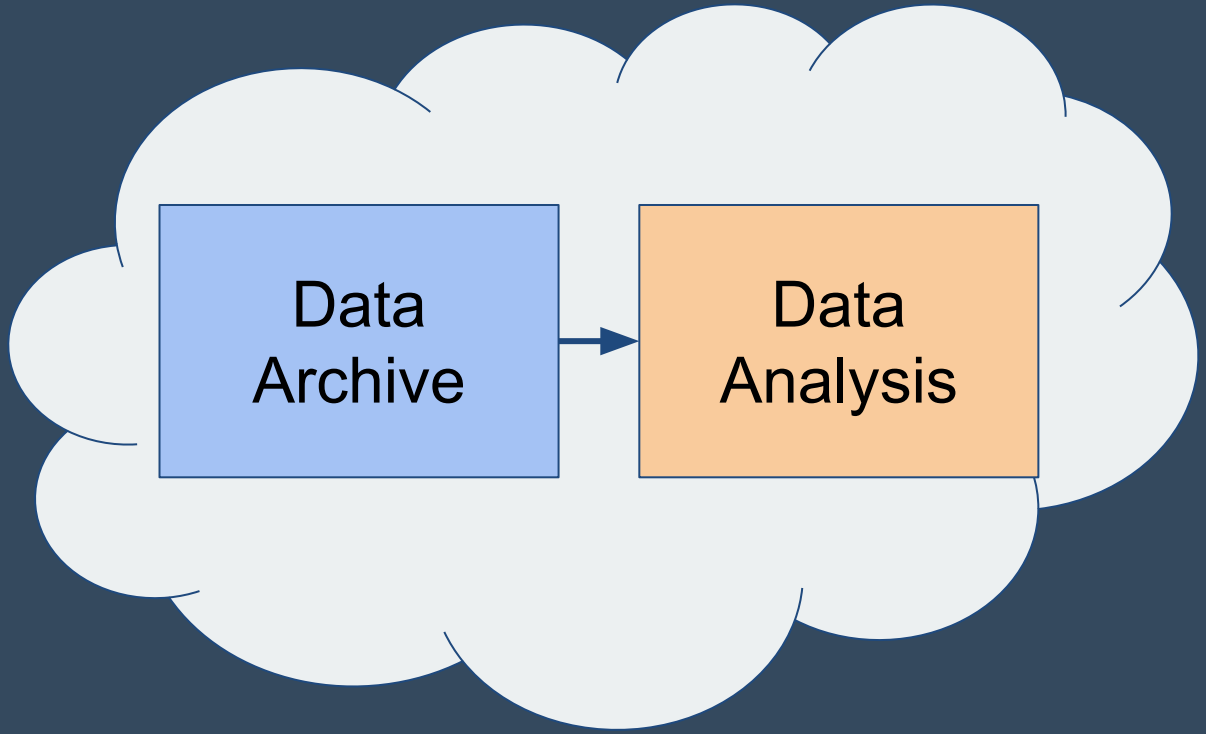
# Projected Data Volumes



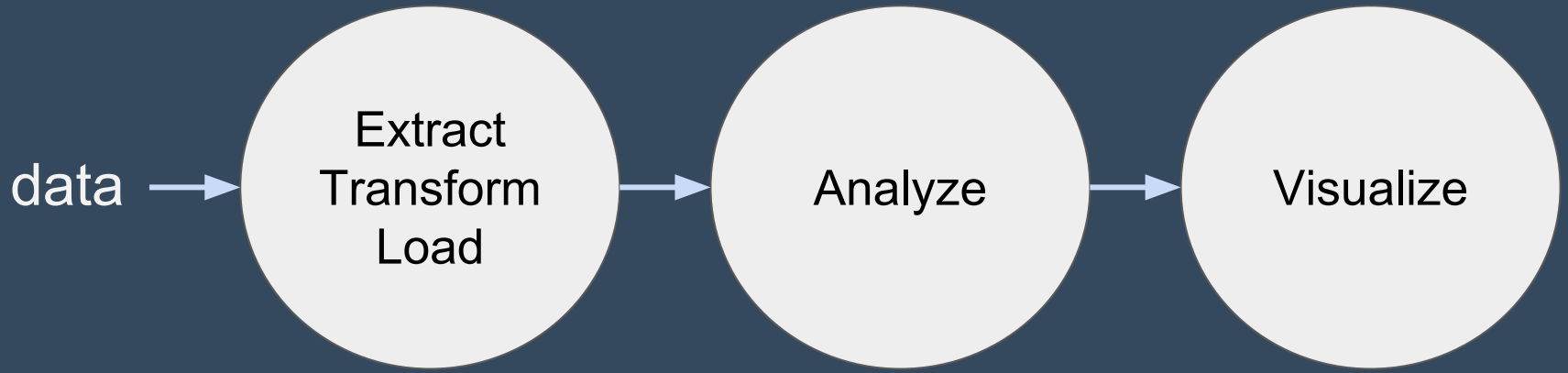
# Distribution increases similarly to cumulative volume



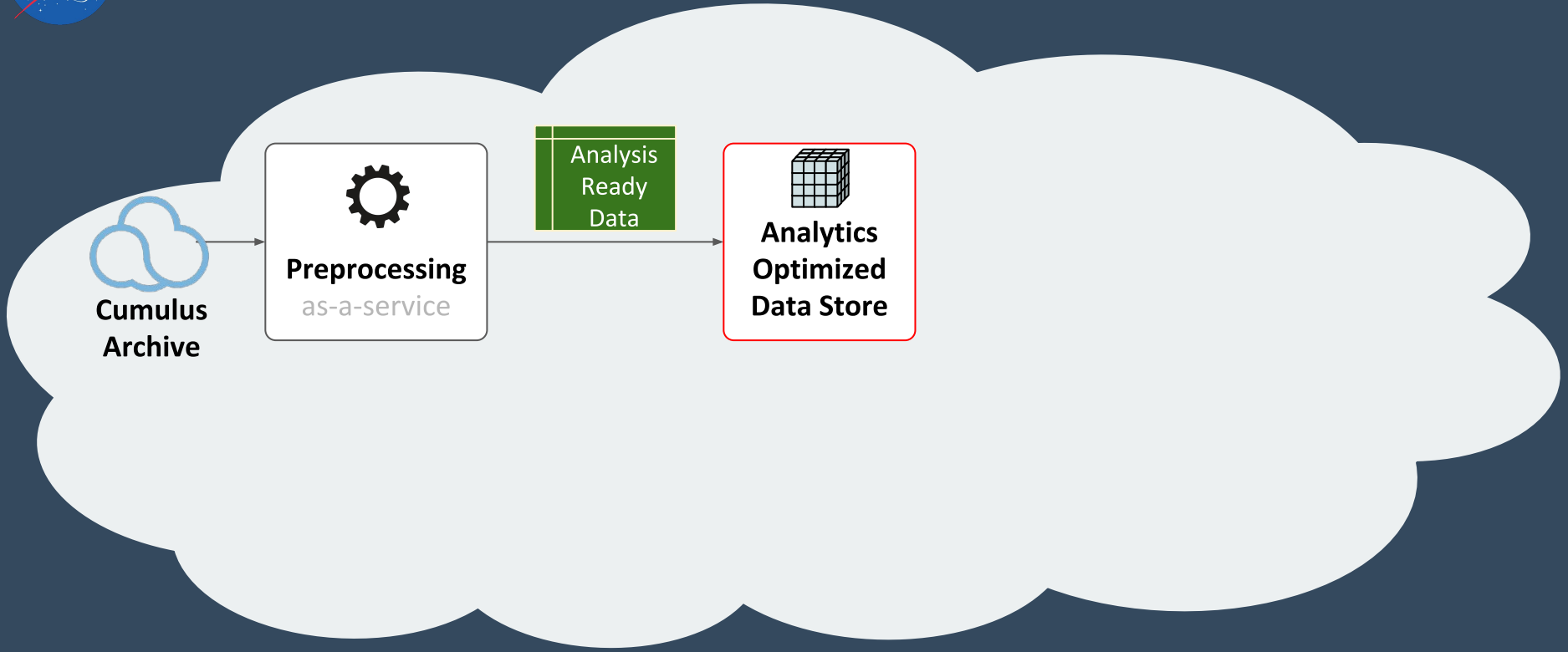
# Solution:  Data-proximal Analysis



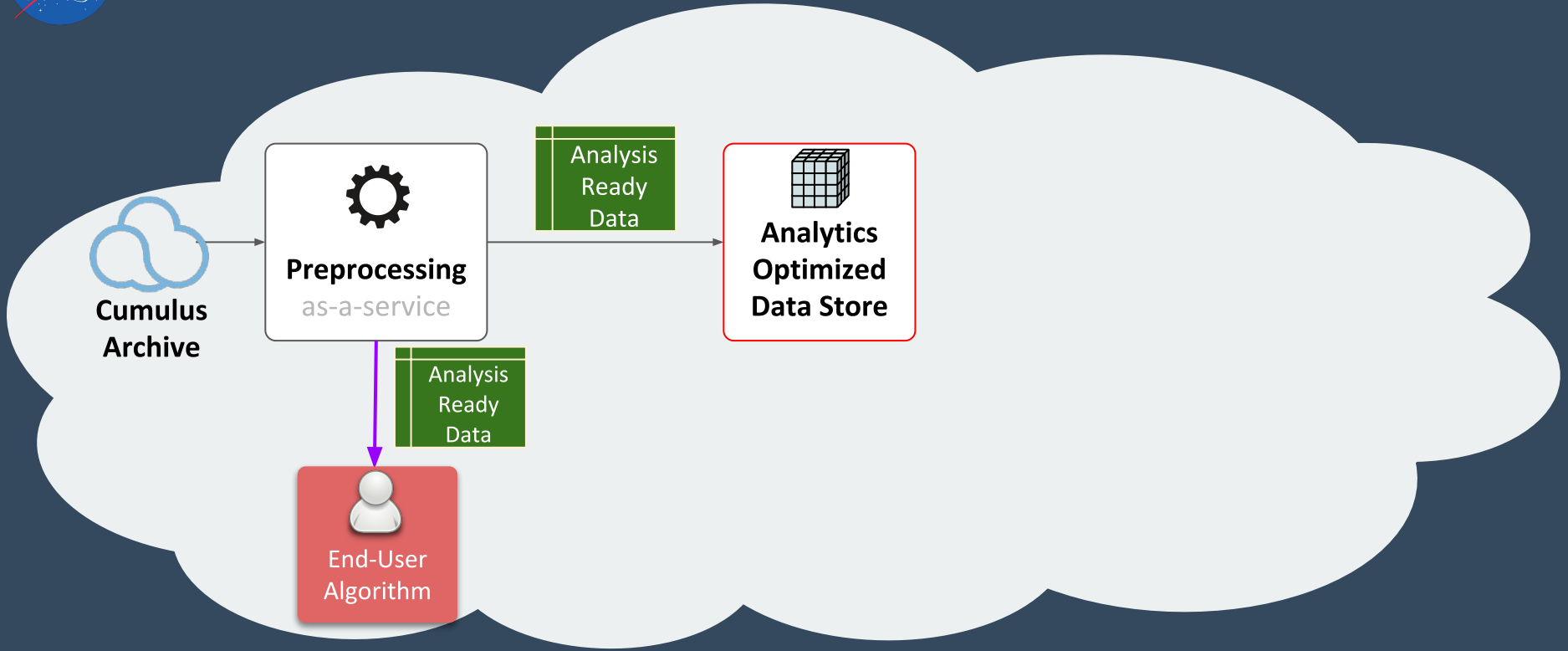
# Abstract Analytics Workflow



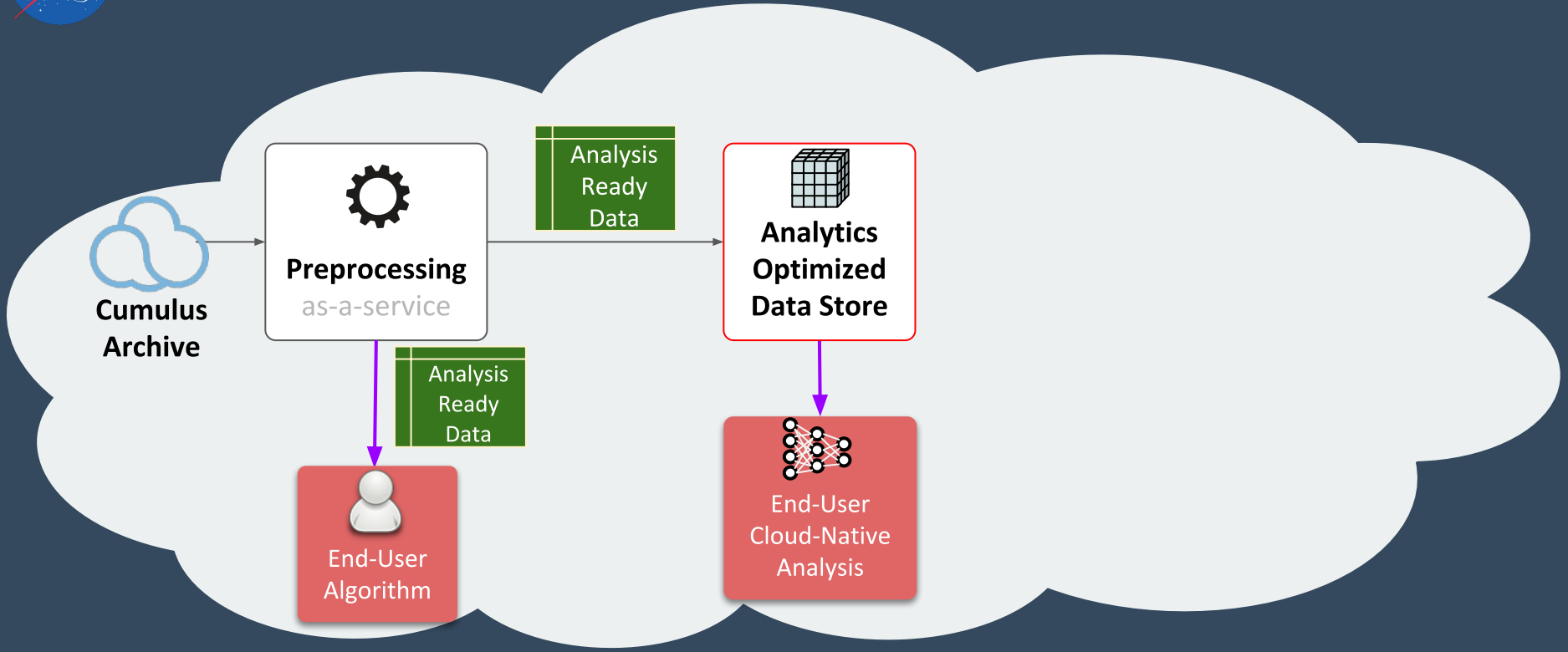
# Extract-and-Transform Preprocessing



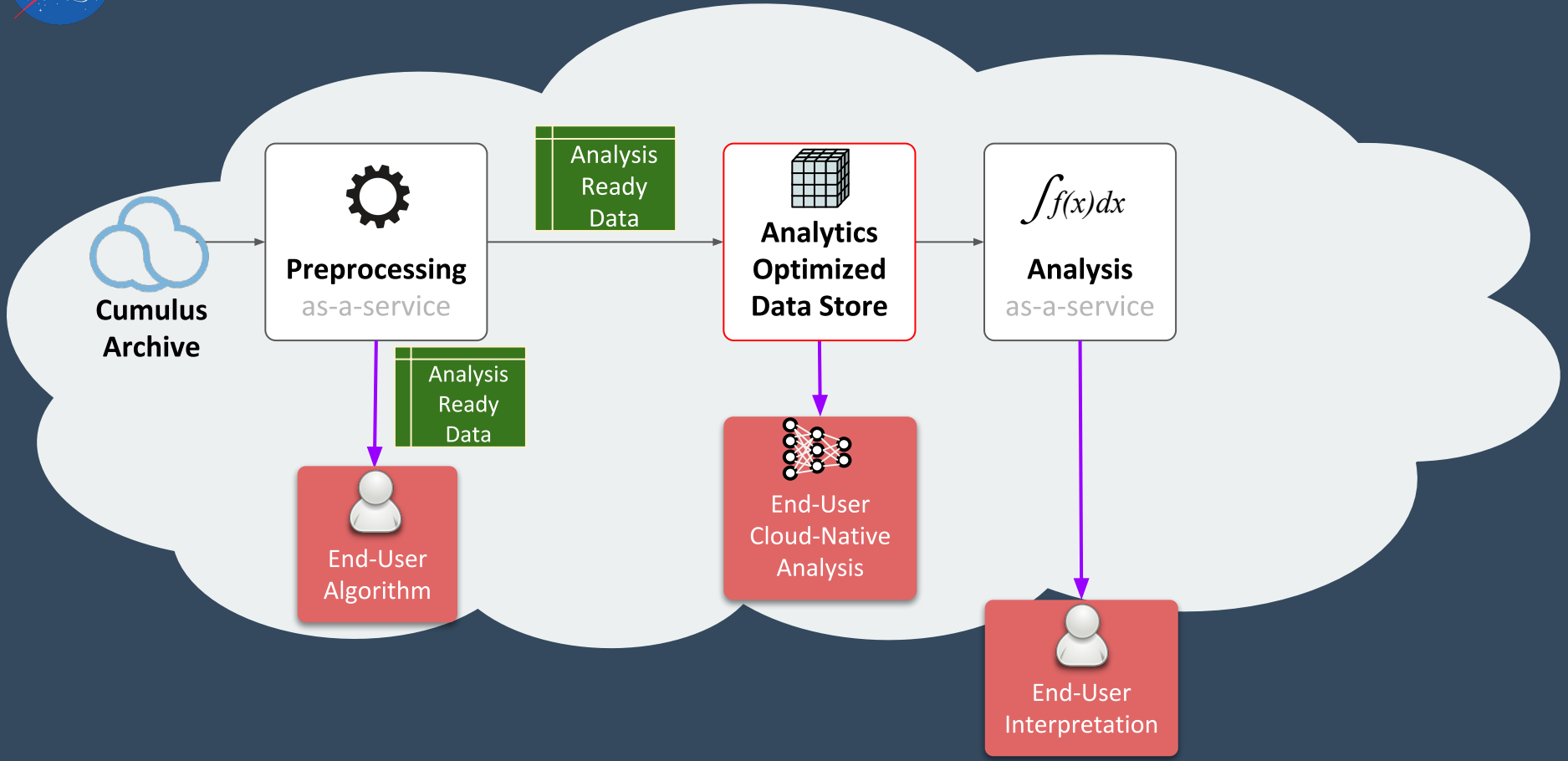
# Analysis-Ready Data for End Users



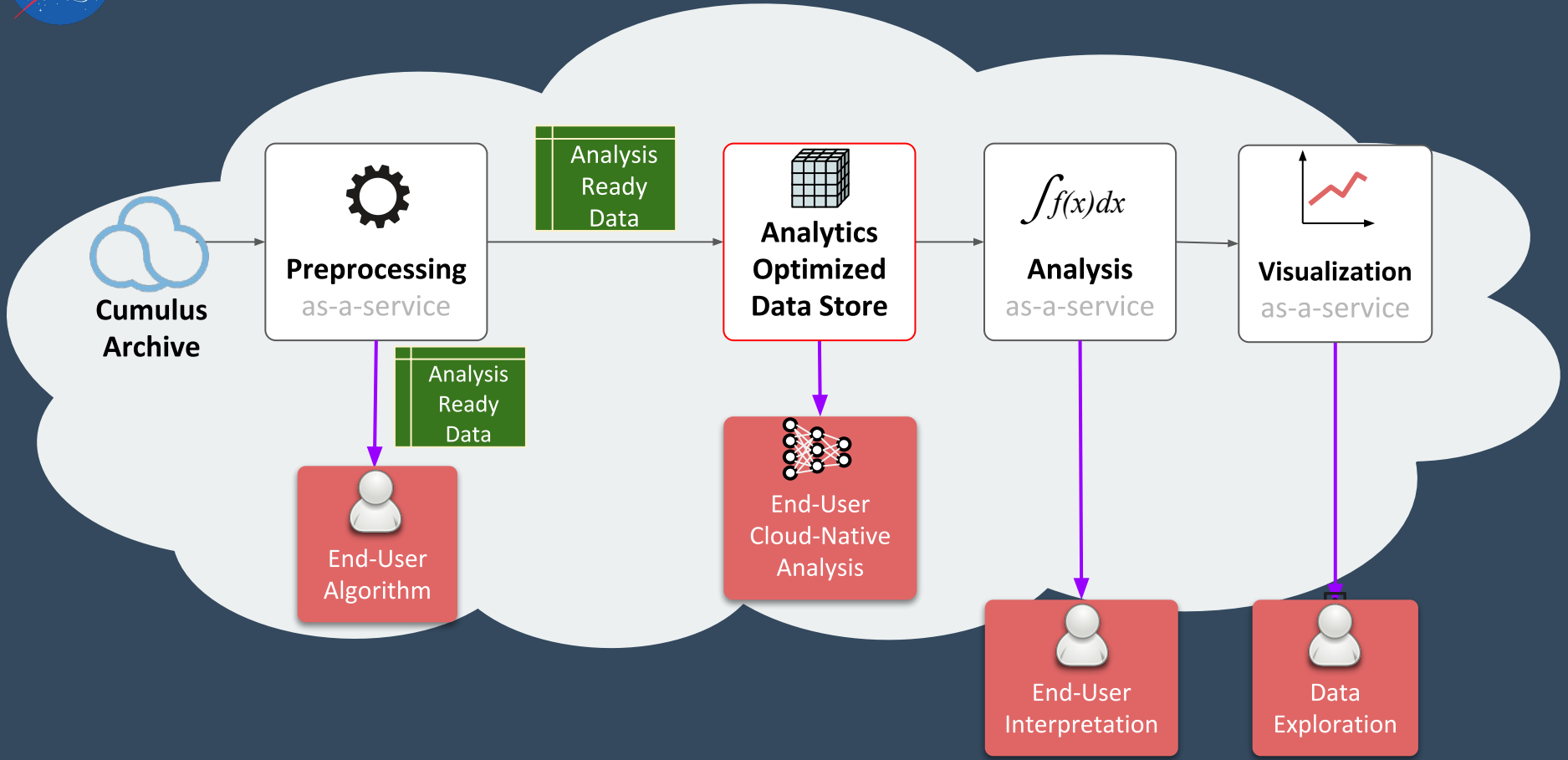
# Cloud-Native Analysis for Data Scientists



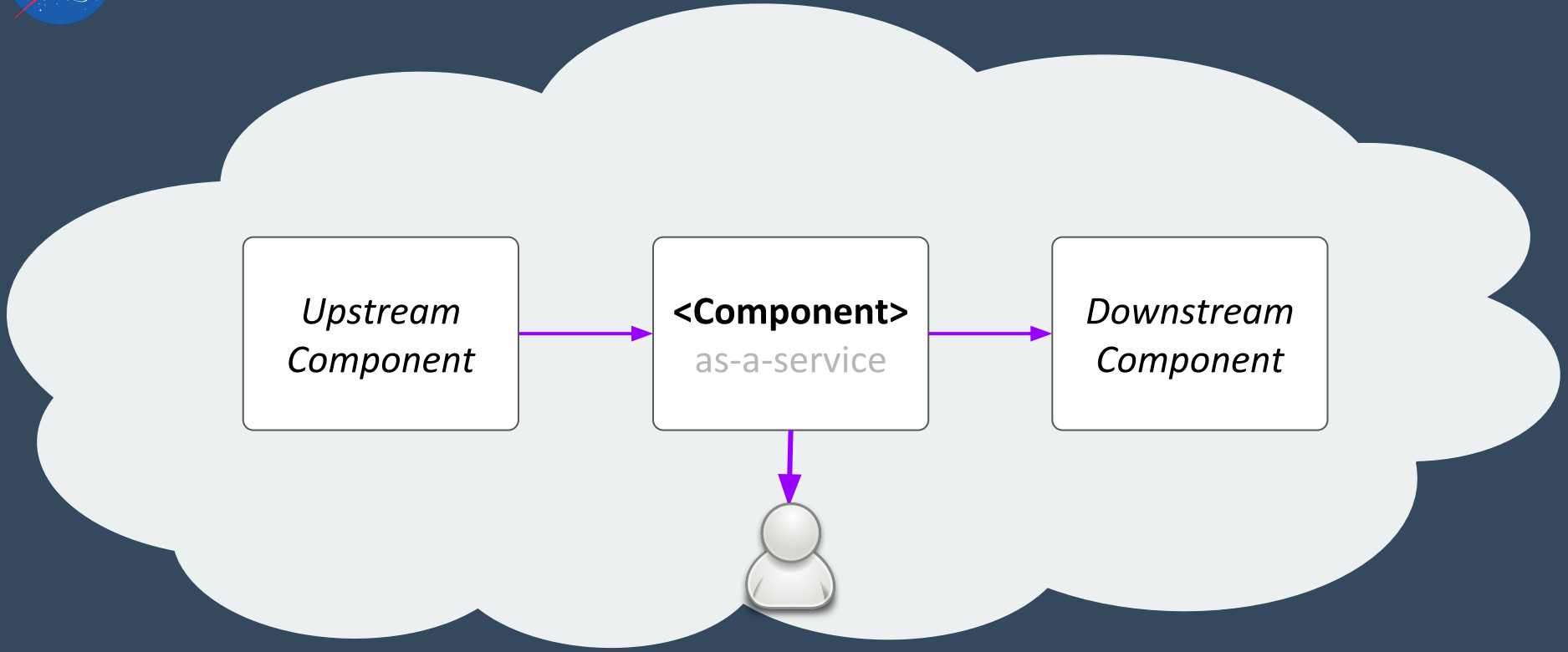
# Analysis-as-a-Service for Interdisciplinary Users



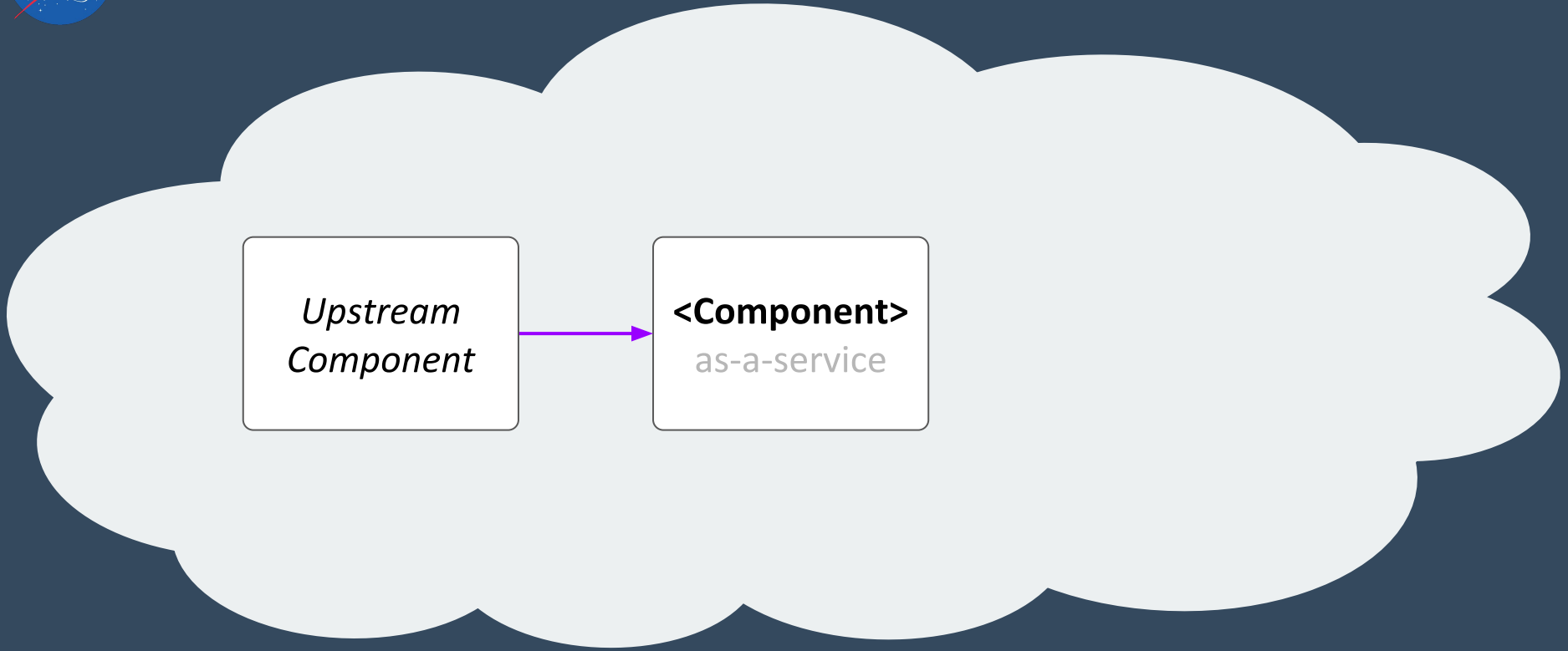
# Interactive Data Exploration (for everybody)



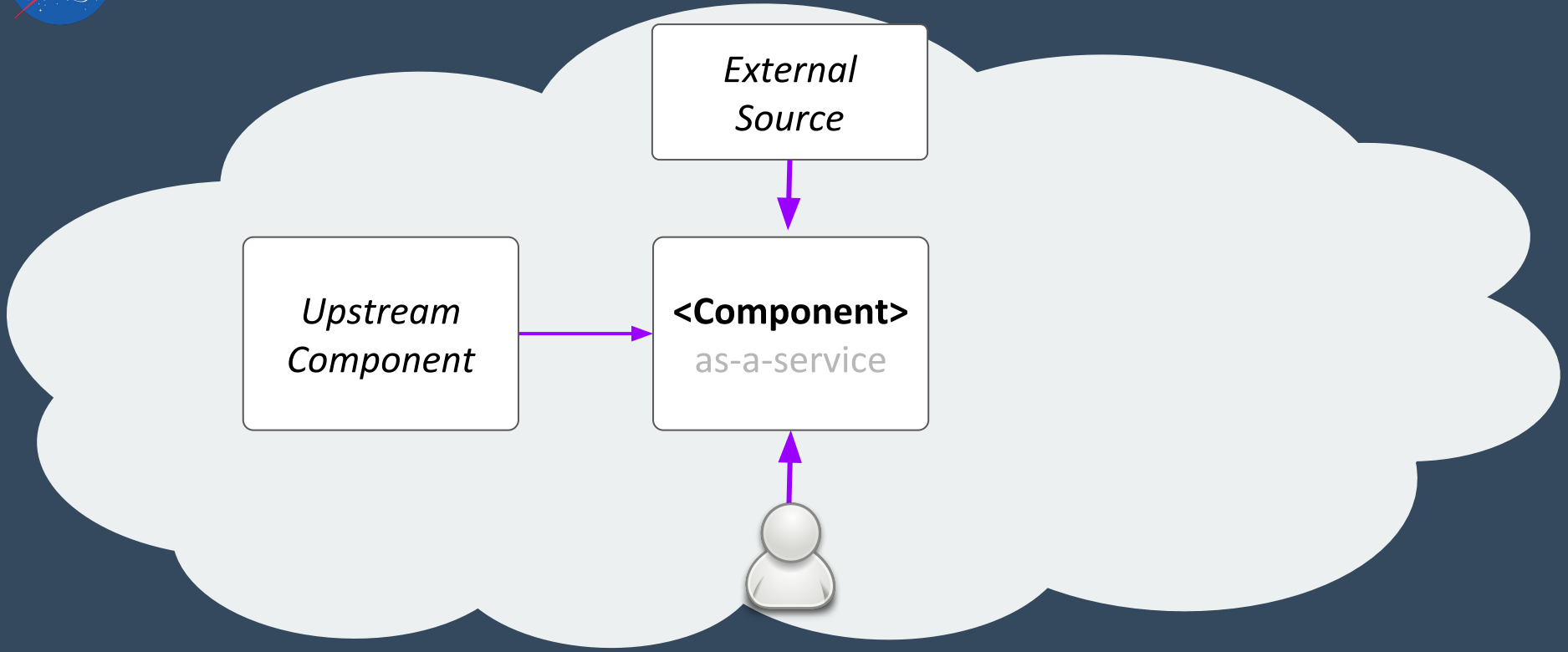
# Service offering enables access throughout the value chain



# Service consumption opens pipeline to external data



# Service consumption opens pipeline to external data



# NASA Earth Science Cloud Analytics Workshop

February 2018, Annapolis, Maryland

~ 40 participants:

- Earth scientists
- Policymakers
- System architects
- Data scientists

Report: https://ntrs.nasa.gov/search.jsp?R=20180002954

# Workshop Recommendations

1. Align Strategies of Cloud Analytic Efforts
2. **Develop Reference Architecture for Cloud Analytics**
3. **Develop Analytics-Optimized Data Stores**
4. Enable Reuse of Cloud Analytics-Related Services
5. Foster Wider Machine Learning Adoption...
6. Foster Cloud Adoption

# Key Questions

*Q: What service specifications should components use?*

1. High-level: Open Geospatial Consortium family?
   a. Web Coverage Processing Service
   b. Web Processing Service
2. Low-level: OpenAPI*?

*API = Application Programming Interface

# Key Questions

*Q: Which data transformations are common enough to include in production of Analysis-Ready Data?*

1. Subsetting
2. Regridding and reprojection
3. Quality filtering?

# Key Questions

*Q: What is the optimal Analytics Optimized Data Store?*

1. Highly scalable database?
2. Hadoop File System?
3. Data Cubes in Web Object Storage?
4. Xarray / zarr?
5. It depends...